IMAGE AUTO-ANNOTATION USING A STATISTICAL MODEL WITH SALIENT REGIONS

Jiayu Tang, Jonathon S. Hare and Paul H. Lewis

Intelligence, Agents, Multimedia Group,
School of Electronics and Computer Science,
University of Southampton, Southampton, SO17 1BJ, United Kingdom
{jt04r, jsh2, phl}@ecs.soton.ac.uk

ABSTRACT

Traditionally, statistical models for image auto-annotation have been coupled with image segmentation. Considering the performance of the current segmentation algorithms, it can be meaningful to avoid a segmentation stage. In this paper, we propose a new approach to image auto-annotation using statistical models. In this approach, segmentation is avoided through the use of salient regions. The use of the statistical model results in an annotation performance which improves upon our previously proposed saliency-based word propagation technique. We also show that the use of salient regions achieves better results than the use of general image regions or segments.

1. INTRODUCTION

Image retrieval would be more straightforward if all the images in the database were semantically annotated. By using standard text query techniques, images could be found in a manner that would meet the different needs of many users. By combining these text-based approaches with visual content search techniques, users could have much more control over the search.

Previously, researchers have tended to use region-based image descriptors for image auto-annotation; Object-shaped regions generated by segmentation algorithms or uniform, usually rectangular, regions have been popular choices. Rectangular regions are a poor choice for image description because they are not robust to a variety of common image transformations, such as rotation. Current segmentation algorithms are not able to perfectly associate segmented regions to the actual objects that are being described. Undoubtedly, segmentation that is conducted by a fallible algorithm will have an adverse effect on the effectiveness of the auto-annotation algorithm.

Two previous approaches to image auto-annotation have been *statistical inference* and *semantic propagation*. Statistical inference is an unsupervised learning method that attempts to learn the association between visual features and keywords by estimating the probability of keywords given regional image features [1, 2, 3, 4, 5]. Propagation is a supervised learning technique that annotates images by directly comparing image similarity at a purely visual level and then propagating keywords based on the most similar images [6, 7].

In this paper, we propose an approach for automatic annotation of images using a statistical model. However, unlike previous approaches this is achieved not by segmenting images but by using salient regions [8].

2. STATISTICAL MODELS FOR IMAGE ANNOTATION

Statistical models try to reveal the association between visual features and keywords by estimating the joint probability distribution of regional image features and keywords, over a set of labelled training images. Given an unlabelled test image, the joint probability of its visual features and each keyword from the vocabulary can be calculated based on the association previously learnt. Some models attempt to annotate image regions [2, 1], whilst others annotate the whole image [3, 4, 5]. The Cross-Media Relevance Model (CMRM) [3], described briefly below, is in the latter class of models.

2.1. The Cross-Media Relevance Model (CMRM)

Following the derivation of Jeon *et al.* [3], the CMRM model can be described as follows. Suppose there exists a training collection T, of labelled images, and a test collection Q, of unlabelled images.

Firstly, each training image is partitioned into shaped or uniform regions. Secondly, visual features, such as colour, shape or texture, are computed for each region. All of these regional features are clustered according to the similarity between them. These clusters, called 'blobs' [1], can be viewed as *visual* words. Each image in the training set can thus be represented as a set of blobs, $B = \{b_1, ..., b_n\}$, together with a set of annotation keywords, $W = \{w_1, ..., w_m\}$. A joint probability distribution, $P(W, B)$, can then be constructed over the training set. In order to perform auto-annotation, the test images are also partitioned into regions, each of which is assigned to the blob that is closest to it. Thus, each test image can also be represented as a set of blobs $B = \{b_1, ..., b_n\}$. The

annotation process for an image is then a matter of finding the words that maximise the conditional probability $P(W|B) = P(W, B)/P(B)$. The joint probability $P(W, B)$ is computed as joint expectation over the space of distributions $P(\cdot|J)$ defined by the training images $J \in T$. Specifically, given a test image $I \in Q$, whose blob representation is $B_I = \{b_{I_1}, ..., b_{I_n}\}$, the following joint probability is computed for each word w from the vocabulary:

$$P(w, b_{I_1}, ..., b_{I_n}) = \sum_{J \in T} P(J)P(w, b_{I_1}, ..., b_{I_n}|J) \ . \quad (1)$$

The CMRM assumes that the events of observing w_i and $b_{I_1}, ..., b_{I_n}$ are mutually independent once an image J is chosen. Therefore, equation (1) becomes:

$$P(w, b_{I_1}, ..., b_{I_n}) = \sum_{J \in T} P(J)P(w|J)\prod_{i=1}^{n} P(b_{I_i}|J) \ . \quad (2)$$

3. REPRESENTING IMAGES USING LOCAL DESCRIPTORS OF SALIENT REGIONS

Salient interest points and regions have been shown to outperform global image descriptors in terms of content-based image retrieval [8, 9] performance. In our algorithm, we select salient regions by using the method proposed by Lowe [10], in which scale-space peaks are detected in a multi-scale difference-of-Gaussian pyramid. In addition, Lowe's SIFT (Scale Invariant Feature Transform) descriptor [10] is used as the feature descriptor. The SIFT descriptor is a three dimensional histogram of gradient location and orientation. The descriptor is constructed in such a way as to make it relatively invariant to small translations of the sampling regions, as might happen in the presence of imaging noise.

Quantisation is applied to the feature vectors to map them from continuous space into discrete space. Specifically, the k-means clustering algorithm, as used in [2, 1], is adopted to cluster the whole set of SIFT descriptors. Each cluster represents a visual term from the visual vocabulary. The feature vectors of each image in the entire data-set are assigned to the closest cluster, enabling each image to be represented by a set of visual terms.

4. HYBRIDISING CMRM WITH A SALIENCY-BASED IMAGE REPRESENTATION

Most current statistical models annotate images by calculating the probability of keywords given the regional feature-vectors. This requires the images to be segmented, into object-shaped regions [2, 1, 3, 4] or uniform regions [5]. However, segmentation algorithms are known to work imperfectly, and uniform regions are intuitively poor choices. That is to say, fallible segmentation potentially compromises the performance of auto-annotation. If the aim is to attach words to

the entire image, instead of image regions, it is possibly beneficial to circumvent the segmentation stage. Yavlinsky *et al.* [11] have shown that the use of global features like colour and texture is promising.

Saliency-based image auto-annotation models [7] have shown some promise. In our previous work [7], a very simple method was proposed; annotations of the top M (1, 2 or 3) training images that best match the test image, in terms of visual similarity, are directly used as the annotations of the test image in question. The problem of this method is that it can not tell which of the annotations is the one most likely to be correct. In other words, it doesn't rank the keywords as statistical models do.

An alternative approach to auto-annotation, explored here, is to use statistical models with saliency, instead of segmentation. The use of a statistical model for annotation allows the keywords to be ranked by their probabilities. We have adopted the CMRM [3] as the statistical model for our experiment and assume that a set of keywords is related to a set of visual terms created from salient regions. Specifically, instead of calculating the joint probability of keywords and image regions (blobs) [3], we calculate the joint probability of keywords and a set of visual terms. As described in section 3, each training image, J, is represented by its saliency-based visual terms $S = \{s_1, ..., s_n\}$ along with its annotations $W = \{w_1, ..., w_n\}$. For each test image, I, the joint probability of each word from the vocabulary and its visual terms, $S_I = \{s_{I_1}, ..., s_{I_n}\}$, is approximated as the expectation over the whole training set, as follows:

$$P(w, S_I) = \sum_{J \in T} P(J)P(w|J)\prod_{i=1}^{n} P(s_{I_i}|J) \ , \quad (3)$$

where, it is assumed that the events of observing w and s_{I_1}, ..., s_{I_n} are mutually independent once a training image J is selected. $P(J)$ is treated uniformly as $1/N_T$, where N_T is the total number of training images. $P(w|J)$ and $P(b|J)$ are estimated by smoothed maximum likelihood, which is derived from [3], as follows:

$$P(w|J) = (1 - \alpha)\frac{\#(w, J)}{|J|} + \alpha\frac{\#(w, T)}{|T|} \ , \quad (4)$$

$$P(s|J) = (1 - \beta)\frac{\#(s, J)}{|J|} + \beta\frac{\#(s, T)}{|T|} \ , \quad (5)$$

where, $\#(w, J)$ denotes the number of times word w occurs in the caption of J, and $\#(w, T)$ denotes the number of times word w occurs in all the captions of images in T. $\#(s, J)$ is the number of times saliency s occurs in J, and $\#(s, T)$ is that of the whole training set. $|J|$ is the aggregate count of all keywords and visual terms in J, and $|T|$ is that of the whole training set. α and β are smoothing parameters obtained by optimising system performance on a held-out portion of the training set.

In the end of the process, all of the words are ranked in the order of possibility of being the correct annotation for the test image in question. The x top-ranking words are chosen as the annotations.

5. RESULTS AND DISCUSSION

Direct comparisons between the saliency-based CMRM approach with the state-of-the-art methods [2, 1, 3, 4, 5] on the Corel image set [2] are not available, because the Corel images at hand are all thumbnail sized. The small image size means most of the images have only between 10 and 20 salient regions which leads to a poor representation of the image content. However, we compare the saliency-based CMRM with the region-based CMRM, as detailed in [3], on the University of Washington Ground Truth Image Database [12].

The Washington data-set contains 697 public-domain images, each of which has between 1 and 13 keywords indicating the image content. On average there are 4.8 keywords per image. After the original keyword labels were processed by correcting mistakes and merging plurals into singular forms [7], the vocabulary consisted of 170 keywords.

Precision and recall, as well as the *normalised score* proposed by Barnard *et al* [1], are used to measure the performance of our salient-based statistical auto-annotation method:

$$Recall = r/n \ , \qquad (6)$$

$$Precision = r/(r + w) \ , \qquad (7)$$

$$E_{NS}^{(model)} = \frac{r}{n} - \frac{w}{N - n} \ , \qquad (8)$$

where, r is the number of correctly predicted words, n is the actual number of words in the test image, w is the number of wrongly predicted words, and N is the number of words in the vocabulary.

5.1. Experimental Results of Auto-annotation by Saliency-based CMRM

We divided the data-set randomly into 3 parts, with 45% as the training set, 5% as the evaluation set and 50% as the test set. The evaluation set is used to estimate the smoothing parameters, α and β, for the CMRM model. Once the parameters are fixed, the training set and the evaluation set are merged to make a new training set, thus resulting in a training set (50%) and test set (50%) of the same size as that used in the previous work [7]. For the saliency-based CMRM, the number of visual terms was set to 3000 as with our previous work [13]. For the region-based CMRM, the optimum was found when the number of blobs was 300.

Figure 1 shows the precision-recall curve of our saliency-based CMRM method, as well as that of the methods reported in [7], namely the LSI (Latent Semantic Indexing) model, the
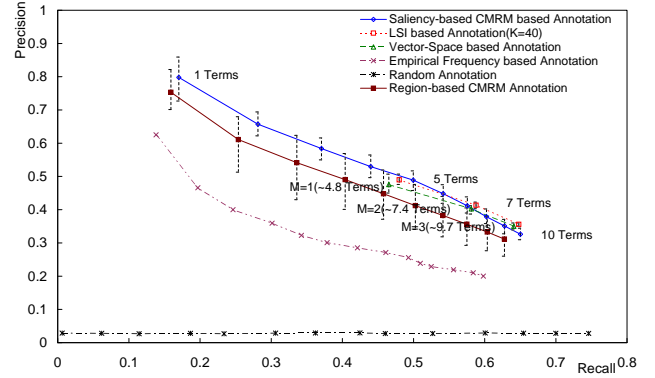

Fig. 1. Precision-Recall curves for several different auto-annotation methods. Error bars show range of precision over 100 repeated runs, each of which used a random separation of the Washington set into training, test and evaluation sets.

Vector Space model, random annotation and empirical frequency based annotation, and the region-based CMRM technique presented in [3]. The curves for the saliency- and region-based CMRM were generated by increasing the number of predicted words from 1 to 10. The results are summarised in Table 1. Figure 2 shows some example images together with their true and predicted annotations.

The results show that auto-annotation using the hybrid CMRM with saliency works much better than by choosing words based on the frequency distribution. Saliency-based CMRM is also capable of predicting words according to the probability of being correct. In the case that only one word for each test image is predicted, up to 80% of predictions are correct. Strictly speaking, this method performs slightly better than the LSI and Vector-Space models when approximately 5 words ($M = 1$) are predicted, but worse for 7 ($M = 2$) and 10 ($M = 3$) predicted keywords. However, accounting for the error bars, these three methods have very similar performances for 5, 7 and 10 words. This implies that the simple annotation propagation methods work almost as well as the statistical method. One possible reason, as argued by Monay and Gatica-Perez [6], could be that propagating annotations can lead to good results when the data-set contains very similar images, which have almost the same set of annotations. This is the case for the Washington Dataset [12]; If the right image is found, the exact annotations are also found. We can also see that on this data-set the saliency-based CMRM performs better than the region-based CMRM.

6. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated a new approach to image auto-annotation by using a statistical model coupled with an image description using salient regions. This approach avoids the image segmentation step taken by many previous auto-annotation techniques. The technique improves on our sim-

Method	Number of Words	Precision	Recall	E_{NS}
Saliency-based CMRM	3	0.584	0.371	0.363
	5	0.489	0.500	0.484
	7	0.412	0.576	0.551
	9	0.351	0.628	0.593
Region-based CMRM	3	0.541	0.336	0.328
	5	0.448	0.458	0.441
	7	0.383	0.541	0.515
	9	0.333	0.604	0.567
Vector-Space	~ 4.8	0.476	0.465	0.450
	~ 7.42	0.402	0.581	0.554
	~ 9.70	0.350	0.641	0.602
LSI(K=40)	~ 4.8	0.490	0.480	0.466
	~ 7.42	0.414	0.588	0.561
	~ 9.70	0.356	0.648	0.609

Table 1. Summary of Results

Images / Methods			
True Annotations	Tree, Bush, Sidewalk	Temple, Sky	Flower, Bush, Tree, Sidewalk, Building
Empirical Annotations	Tree, Building, People, Bush, Grass	Tree, Building, People, Bush, Grass	Tree, Building, People, Bush, Grass
Vector-Space Annotations	Tree, Bush	Tree, Building, Grass, Sidewalk, Pole, People, Clear Sky	Flower, Bush, Tree, Building, Partially Cloudy Sky
LSI Annotations	Tree, Bush, Grass, Sidewalk	Steps, Wall	Flower, Bush, Tree, Ground
Region-based CMRM Annotations	Tree, Flower, Building, Bush, Overcast sky	Tree, Building, People, Clear sky, Cloudy sky	Tree, Building, Bush, Flower, People
Saliency-based CMRM Annotations	Tree, Cloudy sky, Bush, Overcast sky, Post	Clear sky, Rock, Snow, Tree, Building	Tree, Bush, Flower, Ground, Building

Fig. 2. Example Annotations

ple propagation-based annotation methods (LSI and Vector-Space) in the sense that it is able to select individual words. It also improves on the use of general image regions and segments.

Based on this work, we believe that global image descriptors, such as colour histograms, could also be employed within the statistical model used in this paper. More comprehensive comparisons between this technique with other state-of-the-art techniques, such as the CRM model [4], need to be addressed in future work.

7. REFERENCES

[1] K Barnard, P Duygulu, N de Freitas, D Forsyth, D Blei, and M. I Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[2] P Duygulu, K Barnard, J de Freitas, and D Forsyth., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *The Seventh European Conference on Computer Vision*, Copenhagen, Denmark, 2002, pp. IV:97–112.

[3] J Jeon, V Lavrenko, and R Manmatha., "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03*, 2003, pp. 119–126.

[4] V Lavrenko, R Manmatha, and J Jeon., "A model for learning the semantics of pictures," in *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems*, 2003, vol. 16, pp. 553–560.

[5] S. L Feng, R Manmatha, and V Lavrenko., "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of the International Conference on Pattern Recognition (CVPR 2004)*, 2004, vol. 2, pp. 1002–1009.

[6] F Monay and D Gatica-Perez, "On image auto-annotation with latent space models," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 275–278.

[7] J. S Hare and P. H Lewis, "Saliency-based models of image content and their application to auto-annotation by semantic propagation," in *Proceedings of Multimedia and the Semantic Web / European Semantic Web Conference 2005*, 2005.

[8] J. S Hare and P. H Lewis, "Salient regions for query by image content.," in *Image and Video Retrieval: Third International Conference, CIVR 2004*, Dublin, Ireland, July 2004, vol. 3115 of *LNCS*, pp. 317–325, Springer.

[9] N Sebe, Q Tian, E Loupias, M Lew, and T Huang, "Evaluation of salient point techniques," in *CIVR*, 2002, pp. 367–377.

[10] D. G Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] A Yavlinsky, E Schofield, and S. M Rüger, "Automated image annotation using global features and robust non-parametric density estimation.," in *Image and Video Retrieval, 4th International Conference, CIVR 2005*, Singapore, July 2005, LNCS, pp. 507–517, Springer.

[12] University of Washington, "Ground truth image database.," http://www.cs.washington.edu/research /imagedatabase/groundtruth/, 2004.

[13] J. S Hare and P. H Lewis, "On image retrieval using salient regions with vector-spaces and latent semantics.," in *Image and Video Retrieval, 4th International Conference, CIVR 2005*, Singapore, July 2005, vol. 3568 of *LNCS*, pp. 540–549, Springer.